ارزیابی پتانسیل روان‌گرایی خاک در اثر وقوع زمین‌لرزه بااستفاده از چند الگوریتم طبقه‌بندی هوشمند در نرم‌افزار Orange *

هادی فتاحی[1] فاطمه جیریایی[2]

چکیده *یکی از پیامدهای احتمالی وقوع زمین‌لرزه در زمین‌های اشباع، روان‌گرایی خاک و در نتیجهٔ آن شکست و خرابی فونداسیون ساختمان‌ها، انواع زیرساخت‌ها، پل‌ها و بسیاری فجایع دیگر می‌باشد. در این تحقیق سعی شد به‌منظور ارزیابی پتاسیل روان‌گرایی خاک برروی ۷۹ نمونه از پایگاه دادهٔ زلزلهٔ تنگشان کشور چین، چند مدل طبقه‌بندی هوشمند با کمک نرم‌افزار Orange ساخته شود. به همین منظور عملکرد ۵ روش طبقه‌بندی هوشمند (رگرسیون لاجستیک، شبکهٔ عصبی مصنوعی (ANN)، ماشین بردار پشتیبان (SVM)، نزدیک‌ترین همسایگی(KNN) و جنگل تصادفی) براساس معیارهای مختلف با هم مقایسه شدند. نتایج نشان داد روش‌های SVM، ANN و رگرسیون لاجستیک از توانایی بالایی برای پیش‌بینی کلاس روان‌گرایی خاک برخوردار هستند و در بین آنها روش رگرسیون لاجستیک با مقدار شاخص AUC (۰/۹۸) به‌عنوان بهترین روش انتخاب شد. علاوه‌بر این، بررسی تأثیرگذاری متغیرها بااستفاده از چهار معیار بهرهٔ اطلاعاتی، بهرهٔ اطلاعاتی نسبی، شاخص جینی و شاخص ReliefF بیانگر این است که متغیر مقاومت نوک مخروط در آزمایش نفوذ مخروطی مؤثرترین روش است و در اولویت اول قرار می‌گیرد. هم‌چنین متغیرهای نسبت تنش تناوبی و حداکثر شتاب افقی زلزله در سطح زمین ویژگی‌های مهمی به‌حساب می‌آیند.*

واژه‌های کلیدی زمین‌لرزه، روان‌گرایی، الگوریتم‌های طبقه‌بندی هوشمند، نرم‌افزار Orange.

Evaluation of Soil Liquefaction Potential Due to Earthquake using Intelligent Classification Algorithm in Orange Software

H.Fattahi F.Jiryaee

Abstract *One of the possible consequences of earthquakes in saturated areas is soil liquefaction and as a result the failure of foundations of buildings, types of infrastructure, bridges and many other disasters. In this study, in order to evaluate the potential of soil liquefaction on 79 samples from China Tangshan Earthquake Database, several intelligent classification models were constructed with the help of Orange software. Therefore, the performance of 5 intelligent classification methods (Logistic Regression, Artificial Neural Network (ANN), Support Vector Machine (SVM), K-fold Nearest Neighbor (KNN) and Random Forest) were compared based on different criteria. The results showed that SVM, ANN and Logistic Regression methods have a high ability to predict soil liquefaction class and among them the Logistic Regression method with AUC index (0.98) was selected as the best method. In addition, the study of the effectiveness of variables using three criteria of Information Gain, Information Gain Ratio and Gini Index, indicates that the variable measured CPT tip resistance is the most effective variable and is the first priority. The variables of cyclic stress ratio and peak acceleration at the ground surfaceare also important features.*

Key Words: Earthquake, liquefaction, intelligent classification algorithms, Orange software

---

* تاریخ دریافت مقاله۱٤۰۰/۲/۱٦ و تاریخ پذیرش آن ۱٤۰۰/۱۰/۵ از صفحه ۳۹ تا ۵۲ می‌باشد.

(۱) نویسندهٔ مسئول، دانشیار، دانشکدهٔ مهندسی علوم زمین، دانشگاه صنعتی اراک. **Email:** h.fattahi@arakut.ac.ir

(۲) دانشجو، دانشکدهٔ مهندسی علوم زمین، دانشگاه صنعتی اراک.

مقدمه

وجود لایه‌های سست و روان‌گرا و رخداد روان‌گرایی (Liquefaction) در اثر پاسخ لرزه‌ای زمین یکی از مخرب‌ترین حوادث در حوزهٔ ژئوتکنیک می‌باشد. از جمله پیامدهای روان‌گرایی می‌توان به کاهش مقاومت خاک و درنتیجه زمین‌لغزش‌ها، شکست و تخریب فونداسیون‌های ساختمان‌ها و پل‌ها، شناور شدن سازه‌های کم‌وزن مدفون در خاک، جوشش ماسه در سطح زمین، خروج آب از میان ترک‌های سطح زمین و رفتار ماسه‌گونهٔ زمین‌های سخت اشاره کرد [1]. پدیدهٔ روان‌گرایی تنها در خاک‌های اشباع رخ می‌دهد، به‌طوری‌که در مناطق نزدیک آب مانند رودخانه‌ها، دریاچه‌ها، خلیج‌ها و اقیانوس‌ها اثرات تخریبی بیشتری دارد. در محیط‌های اشباع آب در میان ذرات خاک با فشار متعادلی که دارد ذرات خاک را در کنار هم نگه می‌دارد و مانع از حرکت آن‌ها می‌شود. با وقوع لرزش زمین، فشار آب میان منافذ خاک افزایش می‌یابد و ذرات خاک از حالت سکون خارج می‌شوند و در کنار هم شروع به حرکت می‌کنند. در این حالت مقاومت برشی خاک به‌شدت کاهش می‌یابد و به صفر نزدیک می‌شود و موجب روان‌گرایی خاک می‌شود [3,2] . شکل (۱) چگونگی جابه‌جایی ذرات خاک در یک محیط اشباع پس از وقوع زمین‌لرزه را نشان می‌دهد.

ارزیابی پتانسیل روان‌گرایی خاک‌ها اغلب توسط آزمایش‌های صحرایی انجام می‌شود که دو مورد از مرسوم‌ترین آزمایش‌های نفوذی در محل، آزمایش نفوذ استاندارد (Standard Penetration Test) و آزمایش نفوذ مخروطی (Cone Penetration Test) می‌باشد. از طرفی امروزه الگوریتم‌های طبقه‌بندی از مجموعهٔ تکنیک‌های داده‌کاوی توانایی بالایی در پیش‌بینی طبقهٔ متغیرهای چندکلاسه از خود نشان داده‌اندکه بااستفاده از آنها به‌جای روش‌های سنتی می‌توان یک مدل دقیق برای طبقه‌بندی داده‌ها ساخت و سپس از آنها در امر پیش‌بینی استفاده کرد. طبقه‌بندی یکی از شاخه‌های اساسی یادگیری ماشین و داده‌کاوی است و اساس آن داده‌های جمع‌آوری‌شده از اعمال گذشته است [4]. محققان متعددی در این زمینه پژوهش‌هایی انجام داده‌اندکه به این شرح است:

راماکریشنان و همکاران [5] با کمک الگوریتم شبکهٔ عصبی مصنوعی، موضوع روان‌گرایی یک منطقه را مدل‌سازی کردند و مورد مطالعه قرار دادند.

چرن و لی[6] از روشی برمبنای شبکهٔ فازی–عصبی برای ارزیابی پتانسیل روان‌گرایی خاک استفاده کردند و سپس یک روش جستجو برای یافتن نقاط داده‌ها برروی تابع حالت حدی ارائه دادند.

مقیدا [7] در سال ۲۰۰۹ به‌منظور ارزیابی پتانسیل روان‌گرایی خاک، مجموعه داده‌های آزمایش نفوذ مخروطی و طیف گسترده‌ای از پارامترها را با یک برنامهٔ شبکهٔ عصبی مصنوعی تلفیق کرد.

ماریا بااستفاده از برخی از شبکه‌های عصبی مصنوعی به تحلیل شش مسئلهٔ ژئوتکنیکی پرداخت که یکی از این مسائل ارزیابی پتانسیل روان‌گرایی خاک است [8].

سامویی و سیتارام[9] به پیش‌بینی حساسیت روان‌گرایی خاک براساس داده‌های آزمایش نفوذ استاندارد پرداختند و به این منظور از دو روش یادگیری ماشین شبکهٔ عصبی مصنوعی و ماشین بردار پشتیبان استفاده کردند.

فرخزاد و همکاران[10] به زون‌بندی روان‌گرایی خاک در شهر بابل بااستفاده از یک مدل شبکهٔ عصبی پرداختند.

مرت یک مقایسه بین روش‌های آنالیز روان‌گرایی انجام داد و جدولی برای راهنمایی تعیین روش و برنامهٔ صحیح برای ارزیابی پتانسیل روان‌گرایی خاک ارائه داد [11].

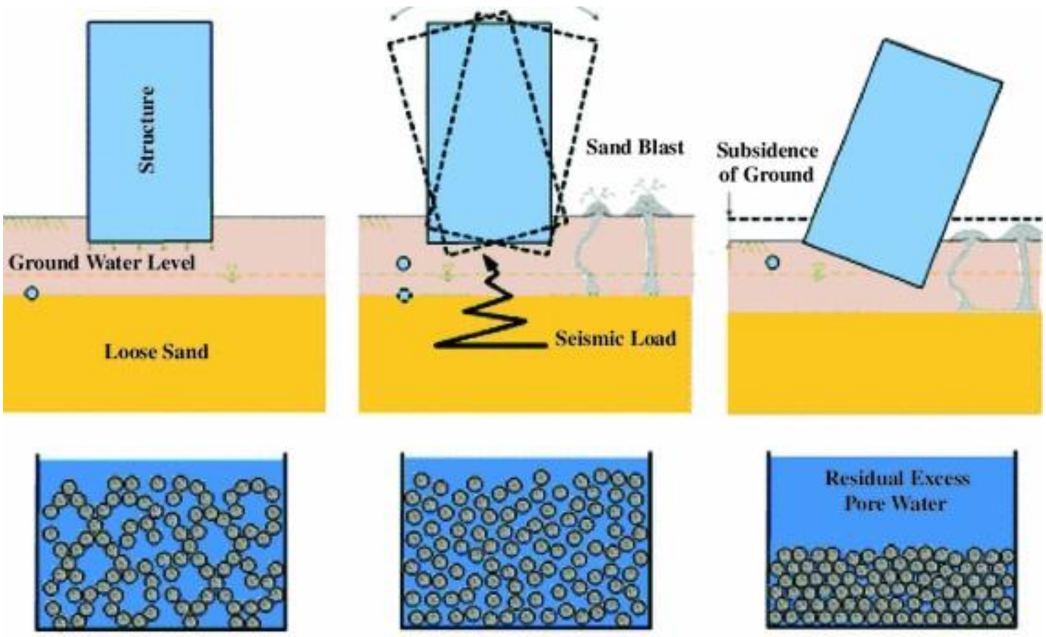
مودولی و داس [12] یک تکنیک تلفیقی هوش مصنوعی و برنامهٔ ژنتیک برای ارزیابی پتانسیل روان‌گرایی خاک براساس داده‌های آزمایش نفوذ استاندارد اجرا کردند.

در این تحقیق به ارزیابی پتانسیل روان‌گرایی خاک پرداخته خواهد شد. این ارزیابی برروی داده‌های میدانی

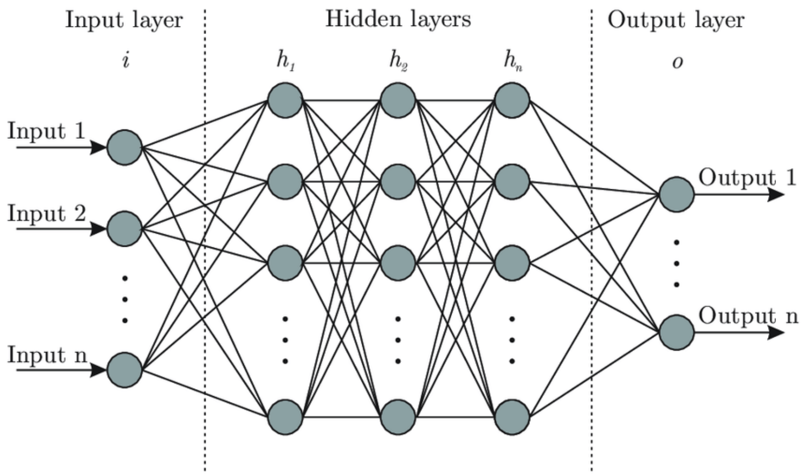
آزمایش نفوذ مخروطی مربوط به زلزلهٔ تانگشان (Tangshan) در کشور چین انجام می‌شود. به این منظور از ۵ مدل طبقه‌بندی هوشمند، رگرسیون لاجستیک، ماشین بردار پشتیبان (SVM)، نزدیک‌ترین همسایگی (KNN) و جنگل تصادفی برای پیش‌بینی کلاس روان‌گرایی خاک استفاده خواهد شد. در این مدل‌ها پارامترهای بزرگی زلزله، سطح آب زیرزمینی، تنش قائم کل، تنش مؤثر قائم، عمق، حداکثر شتاب افقی زلزله در سطح زمین، نسبت تنش تناوبی، میانگین اندازهٔ دانه‌ها، مقاومت نوک مخروط اندازه‌گیری‌شده در آزمایش CPT، به‌عنوان پارامترهای ورودی و پتانسیل روان‌گرایی خاک به‌عنوان پارامتر خروجی می‌باشد و مسئله از نوع طبقه‌بندی است. مدل‌سازی‌ها در نرم‌افزار Orange صورت می‌گیرد. به‌علاوه آنالیز حساسیت پارامترها برای بررسی اهمیت آن‌ها انجام شده‌است.

تکنیک‌های داده‌کاوی در بخش طبقه‌بندی هوشمند

استفاده از ابزار داده‌کاوی برای شناسایی الگوها و مدل‌ها و نیز ارتباط عناصر مختلف در پایگاه داده به‌منظور کشف دانش نهفته در داده‌ها و نهایتا تبدیل داده به اطلاعات، روزبه‌روز ضروری‌تر می‌شود. در سال‌های اخیر استخراج و کشف سریع و دقیق اطلاعات با ارزش و پنهان از مجموعهٔ داده‌ها، به‌عنوان علم داده‌کاوی مورد توجه قرار گرفته‌است که شامل کاربرد ابزارهای مختلف برای تحلیل داده‌های مصنوعی و کشف یک الگوی ناشناختهٔ معتبر بر روابط بین مجموعه داده‌ها می‌باشد. در این بخش به توصیف برخی از تکنیک‌های داده‌کاوی مربوط به طبقه‌بندی پرداخته خواهد شد.



شکل ۱: ساختار ذرات خاک اشباع قبل و پس از وقوع زمین لرزه [۳]



شکل ۲: ساختار بخش‌های یک سلول عصبی [۱۵]

شبکهٔ عصبی مصنوعی (ANN)

می‌توان یک نرون عصبی انسان و عملکرد آن را توسط الگوهای ریاضی الگوسازی کرد. در پردازش اطلاعات یک شبکهٔ عصبی مصنوعی از یک سامانهٔ عصبی زیستی ایده می‌گیرد و مانند مغز به پردازش اطلاعات می‌پردازد. این سامانه از شمار زیادی عناصر پردازشی به نام نرون‌ها تشکیل شده‌است که برای حل یک مسئله با هم هماهنگ عمل می‌کنند. از این‌رو یک شبکهٔ عصبی مصنوعی برای انجام وظیفه‌ای مشخص مانند شناسایی الگوها و دسته‌بندی اطلاعات در طول یک فرآیند یادگیری تنظیم می‌شود. یک شبکهٔ عصبی مصنوعی، از سه لایهٔ ورودی، خروجی و میانی یا پنهان تشکیل می‌شود. ورودی‌ها در وزن‌های مخصوص خود ضرب و با هم جمع می‌شوند و در انتها به‌وسیلهٔ تابع‌هایی خاص خروجی از روی ورودی تصمیم‌گیری می‌شود. نرون می‌تواند یک تابع ریاضی غیرخطی باشد، درنتیجه یک شبکهٔ عصبی که از اجتماع این نرون‌ها تشکیل می‌شود نیز می‌تواند یک سامانهٔ کاملا پیچیده و غیرخطی باشد. یک سلول عصبی از پنج بخش اصلی تشکیل می‌شود که عبارتند از ورودی، وزن‌ها، تابع جمع، تابع فعال‌سازی و خروجی. ورودی‌ها، اطلاعات یا داده‌های خامی هستند که به شبکه تغذیه شده‌است. وزن‌ها مقادیری هستند که اثر یک مجموعهٔ ورودی یا یک عنصر ورودی لایهٔ قبلی را در سلول جدید بیان می‌کنند. تابع جمع، تابعی است که اثر ورودی‌ها و وزن‌ها را به‌طور کامل برروی عنصر موردنظر محاسبه می‌کند. برای استفاده از شبکهٔ عصبی، باید سیستم شبکه ابتدا آموزش ببیند. پس از اتمام آموزش، معمولا خطای شبکه به حداقل می‌رسد و خروجی شبکه نیز مشابه با خروجی هدف خواهد شد [14,13]. شکل (۲) ساختار بخش‌های یک سلول عصبی در یک شبکهٔ عصبی مصنوعی را نشان می‌دهد.

ماشین بردار پشتیبان (SVM)

ماشین بردار پشتیبان، یکی از روش‌های یادگیری با نظارت است که از تئوری یادگیری آماری سرچشمه می‌گیرد و از آن برای طبقه‌بندی و رگرسیون استفاده می‌کنند. SVM پیشگویی‌های خود را بااستفاده از ترکیبی خطی از تابع کرنل که برروی مجموعه‌ای از داده‌های آموزشی با نام بردارهای پشتیبان عمل می‌کند، انجام می‌دهد. در تقسیم‌بندی خطی داده‌ها، سعی می‌شود خطی انتخاب شود که حاشیهٔ اطمینان بیشتری داشته باشد. یکی از خصوصیات مهم بردار پشتیبان این است که به‌طور هم‌زمان خطای تجربی طبقه‌بندی را کمینه و حاشیه‌های هندسی را بیشینه می‌کند؛ بنابراین طبقه‌بندی بیشینه‌کنندهٔ حاشیه نیز نامیده می‌شود. در ماشین بردار پشتیبان، هدف به حداکثر رساندن حاشیهٔ بین دو کلاس است؛ بنابراین ابرصفحه‌ای انتخاب می‌شود که فاصلهٔ آن از نزدیک‌ترین داده‌ها در هر دو طرف جداکنندهٔ خطی، بیشینه باشد. این ابرصفحه از طریق رابطهٔ زیر به‌دست می‌آید [15].

$$W^T \emptyset(x) + b = 0 \quad (۱)$$

در این رابطه بردار وزن w برداری عمود بر ابرصفحه است. b بردار بایاس است که به‌منظور اندازه‌گیری فاصلهٔ ابرصفحه تا مبدأ استفاده می‌شود. $\phi(x)$ کرنلی برای انتقال داده به فضایی با ابعاد بالاتر است.

نزدیک‌ترین همسایگی (KNN)

به‌طور کلی الگوریتم نزدیک‌ترین همسایگی دوکاربرد دارد. کاربرد اول تخمین تابع چگالی توزیع داده‌های تعلیم و کاربرد دوم طبقه‌بندی داده‌های تست می‌باشد. اساس کار این روش تخمین ویژگی‌های یک سری داده‌های مجهول باتوجه به بیشترین شباهت این داده‌ها با داده‌های معلوم که در همسایگی یا نزدیکی آنها قرار دارند، می‌باشد. گام اول در این الگوریتم انتخاب روش یا رابطه‌ای برای محاسبهٔ فاصلهٔ بین داده‌های مورد آزمایش و داده‌های تعلیم می‌باشد که در اکثر موارد از فاصلهٔ اقلیدسی در رابطهٔ زیر استفاده می‌شود [16].

$$d(p.q) = \sqrt{\sum_{i=1}^{n} (p_i - q_i)^2} \quad (۲)$$

n تعداد فیلدهای هر رکود است و p_i و q_i مقادیر ویژگی iام برای رکودها هستند. ابتدا فاصلهٔ رکورد جدید از همهٔ رکوردهای آموزشی محاسبه می‌شود. سپس K نزدیک‌ترین رکوردها را به نمونهٔ جدید براساس یک معیار شباهت به‌دست می‌آورد و دسته‌های این K همسایه را بررسی می‌کند. در آخر، دستهٔ نمونهٔ جدید را برابر با بیشترین دسته در K همسایهٔ آن قرار می‌دهد.

رگرسیون لاجستیک

در تحلیل مسائل چندمتغیره در حالتی‌که متغیر هدف به‌صورت متغیر دودویی باشد از مدل رگرسیون لاجستیک برای طبقه‌بندی داده‌ها استفاده می‌شود. در این مدل رابطهٔ رگرسیونی متغیرها خطی نیست و به‌صورت منحنی S شکل است. در این روش پس از محاسبهٔ احتمال عضویت نمونه، مقدار احتمال بیشتر از ۰/۵ به ۱ و کمتر از ۰/۵ به صفر تبدیل می‌شود. رابطهٔ رگرسیون لاجستیک به شکل زیر است [17].

$$P(Y_i = 1|X_i) = \frac{\exp(B_1 + B_2X_i + \cdots + B_nX_i)}{1 + \exp(B_1 + B_2X_i + \cdots + B_nX_i)} \quad (۳)$$

در این رابطه P احتمال وقوع، B_1 ضریب ثابت، B_n ضریب زاویهٔ متغیرها یا عرض از مبدأ، Y_i متغیر وابسته و X_i متغیر مستقل است.

جنگل تصادفی

روش جنگل تصادفی یک نوع مدرن از روش‌های درخت-پایه است که از تعداد زیادی درخت‌های کلاس‌بندی و رگرسیونی تشکیل شده‌است. هم‌چنین یکی از روش‌های ناپارامتریک مناسب برای مدل‌سازی داده‌های پیوسته و گسسته روش درخت تصمیم می‌باشد. جنگل تصادفی بااستفاده از مجموعه‌ای از درخت‌ها با در نظر گرفتن n داده مشاهدهٔ مستقل ساخته می‌شود، به‌طوری‌که با ترکیب چندین درخت تصمیم دقت مدل را بالا می‌برد. هر درخت تصمیم‌گیری بااستفاده از یک نمونهٔ تصادفی از مجموعه نمونه‌های تعلیم، آموزش می‌بیند. هم‌چنین انتخاب متغیرهای پیش‌بینی‌کننده برای تقسیم‌بندی گره‌ها به‌صورت تصادفی انجام می‌شود. در روش جنگل تصادفی ویژگی mtry برای تعداد متغیرهای کمکی مورد استفاده در هر زیرمجموعه و یا هر گره درخت تصمیم و ntree تعداد درختان مورد استفاده در جنگل تصادفی است که به‌عنوان پارامترهایی برای این روش بایستی تنظیم شود. یکی از قابلیت‌های این روش تخمین اهمیت و تأثیر متغیرهای کمکی بااستفاده از تغییر خطا در صورت وجود و عدم وجود آن متغیر است [18].

روش‌های ارزش‌گذاری (Scoring Method) به‌کار گرفت شده

الگوریتم درخت تصمیم از جمله روش‌های طبقه‌بندی و رگرسیون است که انواع مختلفی دارد. یک موضوع مهم در ساخت این درخت‌ها انتخاب

شایسته‌ترین صفت‌ها (Attribute Selection Method) یا ویژگی‌ها در سطوح بالاتر یا نزدیک به ریشه است و در هر نوع از درخت تصمیم می‌توان از روش‌های مختلف ارزش‌گذاری یا روش‌های انتخاب صفت کمک گرفت. یک نوع از درخت تصمیم ID3 است که از روش بهرهٔ اطلاعاتی (Information Gain) استفاده می‌کند. هرچه مقدار این شاخص برای یک ویژگی بالاتر باشد، اطلاعات بیشتری توسط آن ویژگی گرفته می‌شود و بهتر می‌توان مجموعهٔ داده‌ها را براساس آن ویژگی کلاس‌بندی کرد. نوع دیگر درخت تصمیم CART (Classification and Regression Tree) است که براساس متغیرهای دودویی بنا نهاده شده‌است و از معیاری به نام شاخص جینی (Gini Index) برای انتخاب صفت‌ها کمک می‌گیرد. هر چه شاخص جینی کمتر باشد یعنی آن ویژگی اطلاعات بیشتری به ما می‌دهد و می‌تواند در درخت ساخته‌شده در سطوح بالاتر و نزدیک به ریشه قرار بگیرد. معیار دیگری به نام بهرهٔ اطلاعاتی نسبی (Informatio Gain Ratio) وجود دارد که بهتر از شاخص بهرهٔ اطلاعاتی عمل می‌کند. در بهرهٔ اطلاعاتی نسبی از بین ویژگی‌ها، آن‌که نسبت بهرهٔ اطلاعاتی به آنتروپی آن بزرگ‌تر باشد وزن بیشتری خواهد داشت. معیار بهرهٔ اطلاعاتی خود از معیار آنتروپی استفاده می‌کند. روابط (٤) تا (٦)، چگونگی محاسبهٔ بهرهٔ اطلاعاتی نسبی را بیان می‌کند [19].

$$Information\ Gain\ (A) = Entropy\ (D) - Entropy_A(D) \quad (٤)$$

رابطهٔ (٤) بهرهٔ اطلاعاتی را برای پارامتر (A) محاسبه می‌کند که در آن D دلالت بر مجموعه داده‌های آموزشی دارد:

$$Entropy\ (D) = -\sum_{i=1}^{c} P_i \times Log_2(P_i) \quad (٥)$$

$$Entropy_A\ (D) = \sum_{j=1}^{v} \frac{|D_j|}{|D|} \times Entropy(D_j) \quad (٦)$$

که در آن c تعداد برچسب کلاس‌های موجود در داده‌های آموزشی، P_i احتمال این‌که نمونه‌ای از داده‌ها متعلق به کلاس iام باشد، v تعداد اعضای دامنهٔ پارامتر A و D_j قسمتی از داده‌های اولیه که مقدار پارامتر آن‌ها v_j است را نشان می‌دهد. در ضمن $|D|$ دلالت بر اندازهٔ داده‌های D دارد.

علاوه‌بر معیارهای نام‌برده شده، روش دیگری که برای آنالیز حساسیت یا به‌عبارتی ارزیابی اهمیت پارامترهای ورودی به‌کار گرفته می‌شود استفاده از معیار ReliefF است. ReliefF الگوریتمی است که توسط Kira و Rendell در سال ۱۹۹۲ توسعه یافته‌است [20]. این معیار رویکردی با روش فیلتر برای انتخاب پارامترها دارد و به‌طور قابل توجهی به روابط بین پارامترها حساسیت نشان می‌دهد. این روش امتیازی برای هر پارامتر محاسبه می‌کند که می‌تواند برای رتبه‌بندی و انتخاب پارامترهای برتر استفاده شود .

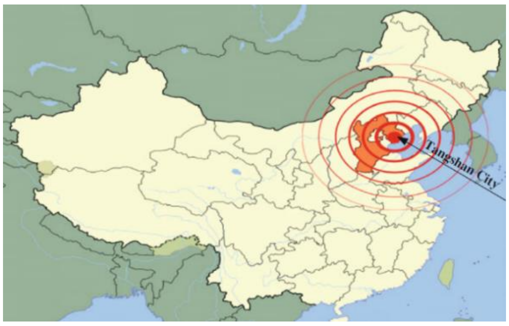
معرفی نرم‌افزار Orange

نرم‌افزار Orange یک ابزار داده‌کاوی بسیار کارآمد برپایهٔ زبان برنامه‌نویسی پایتون است که بااستفاده از آن به‌صورت تعاملی و کاملا بصری می‌توان عملیات داده‌کاوی را بدون نیاز به کدنویسی انجام داد و خروجی مناسبی را تهیه نمود. کار با این نرم‌افزار به‌صورت گرافیکی است؛ به همین دلیل مدل‌سازی‌ها را بسیار آسان و قابل فهم کرده‌است؛ به‌طوری‌که به‌صورت هم‌زمان امکان کشف اطلاعات از چندین مجموعه داده در طی مدل‌سازی‌های مختلف را می‌دهد. این نرم‌افزار هم‌چنین دارای ویژگی متن باز (Open source) است و قابلیت اضافه کردن کدهای دل‌خواه و تکمیل مدل‌های موجود را دارد. نرم‌افزار Orange شامل مجموعه‌ای از الگوریتم‌های یادگیری ماشین تحت نظارت برای طبقه‌بندی و رگرسیون، روش‌های اعتبارسنجی براساس

نمونه‌برداری و ارزیابی قابل اطمینان، الگوریتم‌های بدون نظارت یادگیری برای خوشه‌بندی، الگوریتم‌های قواعد انجمنی، الگوریتم‌هایی برای پردازش زبان طبیعی و استخراج متن و الگوریتم‌هایی برای تجزیه و تحلیل سری‌های زمانی و مدل‌سازی و ... است.

تحلیل و نتایج

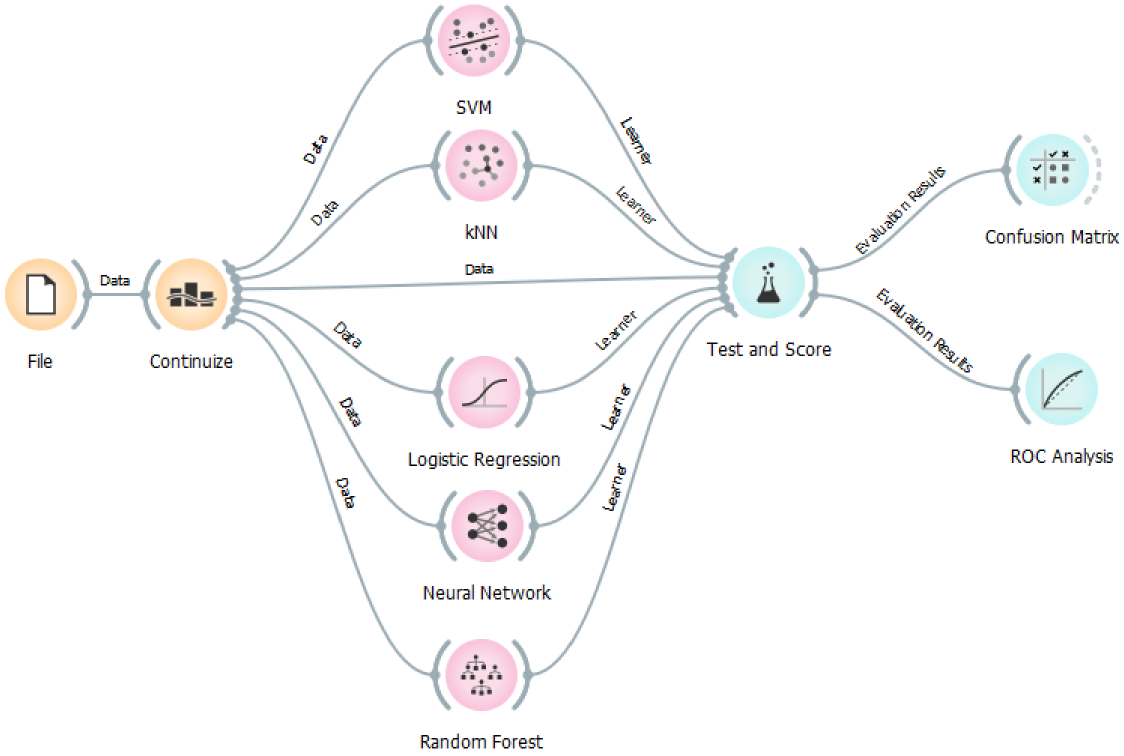
ارزیابی پتانسیل روان‌گرایی خاک یکی از مسائل پیچیدۀ ژئوتکنیک است که بررسی آن با تأثیر گرفتن از چندین فاکتور اعم از خصوصیات خاک، شرایط زمین‌شناسی و ویژگی‌های حرکت زمین انجام می‌شود. در این قسمت به‌منظور پیش‌بینی پتانسیل روان‌گرایی خاک در یک پایگاه داده، از نرم‌افزار Orange کمک گرفته می‌شود و با به‌کار بستن روش‌های مختلف مدل‌سازی، کارایی و عملکرد آنها در یک پیش‌بینی دقیق و صحیح مقایسه خواهد شد. این پایگاه داده مربوط به زلزلۀ تانگشان به بزرگی ۷/۸ ریشتر است، یک فاجعۀ طبیعی که در روز چهارشنبه ۲۸ ژوئیۀ ۱۹۷۶ در کشور چین رخ داده‌است. مرکز زمین‌لرزه در نزدیکی تانگشان در هبی (Hebei)، جمهوری خلق چین، یک شهر صنعتی با حدود یک میلیون نفر است (شکل ۳) [3]. مشخصات آماری این پایگاه داده در جدول (۱) آورده شده‌است. پتانسیل روان‌گرایی خاک به‌عنوان متغیر هدف درنظر گرفته می‌شود و از نوع دودویی با دو مقدار صفر، یعنی عدم رخداد روان‌گرایی و ۱ یعنی رخداد پدیدۀ روان‌گرایی می‌باشد. همچنین ورودی مدل‌ها شامل ۱ متغیر قطعی و ۸ متغیر تصادفی است که عبارتند از بزرگی زلزله (M)، سطح آب زیرزمینی (d_w)، تنش قائم کل (σ_v)، تنش مؤثر قائم (σ'_{v0})، عمق (d_s)، حداکثر شتاب افقی زلزله در سطح زمین (α_{max})، نسبت تنش تناوبی (τ_{av}/σ'_{v0})، میانگین اندازۀ دانه‌ها (D_{50})، مقاومت نوک مخروط اندازه‌گیری‌شده در آزمایش CPT (q_c). نمونه‌ها شامل ۷۹ دادۀ میدانی روان‌گرایی خاک براساس آزمون نفوذ مخروطی هستند. داده‌ها قبل از ورود به مدل بین مقدار ۱- و ۱ بدون بعد می‌شوند. شمای کلی مدل‌سازی انجام‌شده در نرم‌افزار در شکل (۴) نشان داده شده‌است. در روند مدل‌سازی همان‌طور که مشخص است، در ابتدا چیزی که مهم است نرمال کردن داده‌های ورودی است. سپس این داده‌های تبدیل یافته به ۵ مدل طبقه‌بندی از الگوریتم‌های موجود وارد می‌شوند و مدل‌های KNN، SVM، ANN، جنگل تصادفی و رگرسیون لاجستیک ساخته می‌شوند. درنهایت برای ارزیابی عملکرد آنها از امکانات موجود مانند ماتریس درهم‌ریختگی (Confusion Matrix) و نمودار ROC (Receiver Operating Characteristic) کمک گرفته شده‌است.



شکل ۳: نمایی از مکان زلزله تنگشان [۳]

جدول ۲: مقدار شاخص‌های ارزیابی روش‌های طبقه بندی به‌ازای هر روش

مدل‌های طبقه بندی	مساحت زیر منحنی ROC	صحت طبقه‌بندی	F_1	دقت طبقه‌بندی	بازیابی مدل
نزدیک‌ترین همسایگی	۰/۹۳۹	۰/۹۲۴	۰/۹۲۳	۰/۹۲۴	۰/۹۲۴
ماشین بردار پشتیبان	۰/۹۱۴	۰/۹۳۷	۰/۹۳۴	۰/۹۴۲	۰/۹۳۷
جنگل تصادفی	۰/۹۳۸	۰/۸۸۶	۰/۸۸۲	۰/۸۸۷	۰/۸۸۶
شبکه عصبی هوشمند	۰/۹۱۴	۰/۹۳۷	۰/۹۳۵	۰/۹۳۸	۰/۹۳۷
رگرسیون لاجستیک	۰/۹۷۹	۰/۹۳۷	۰/۹۳۶	۰/۹۳۶	۰/۹۳۷



شکل ٤: شمای کلی مدلسازی انجام شده در نرم‌افزار Orange

در جدول (۲) برای هر مدل طبقه‌بندی مقدار شاخص‌های مختلفی محاسبه شده‌است و براساس آنها می‌توان به مقایسهٔ عملکرد مدل‌ها پرداخت. همهٔ شاخص‌های به‌کاربرده شده از طریق ماتریس درهم‌ریختگی به‌دست آمده‌اند. در این جدول AUC (Receiver Operating Characteristic)، مساحت زیر منحنی ROC، CA نسبت نمونه‌های درست پیش‌بینی‌شده به تعداد کل نمونه‌ها، Precision نسبت نمونه‌های مثبت صحیح به تعداد نمونه‌هایی که مثبت پیش‌بینی شده‌اند، Recall نسبت نمونه‌های مثبت صحیح به تعداد نمونه‌هایی که در واقعیت مثبت بوده‌اند و F_1 یک میانگین هارمونیک وزن‌دار شدهٔ دو شاخص Recall و Precision می‌باشد که در رابطهٔ (۷) بیان شده‌است.

$$F_1 = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (۷)$$

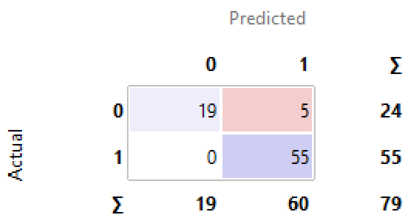
شکل (۵) ماتریس درهم‌ریختگی مدل ماشین بردار پشتیبان را نشان می‌دهد. باتوجه به تعاریفی که از معیارها بیان شد، مقدار هر معیار در مدل SVM طبق رابطهٔ (۸) تا (۱۱) محاسبه می‌شود. بایستی دقت کرد که درمورد دو معیار دقت و بازیابی، در هرکلاس به‌طور مجزا محاسبه می‌شود و از آنها میانگین وزن‌دار گرفته می‌شود. وزن هر کلاس متناسب با نسبت نمونه‌های آن کلاس به تعداد کل نمونه‌ها است.

$$CA = \frac{19 + 55}{79} = 0\cdot9367 \quad (۸)$$

$$\text{Precision} = \left(\frac{24}{79} \times \frac{19}{19}\right) + \left(\frac{55}{79} \times \frac{55}{60}\right) = 0\cdot9419 \quad (۹)$$

$$\text{Recall} = \left(\frac{24}{79} \times \frac{19}{24}\right) + \left(\frac{55}{79} \times \frac{55}{55}\right) = 0\cdot9367 \quad (۱۰)$$

$$F_1 = \frac{2 \times 0\cdot9419 \times 0\cdot9367}{0\cdot9419 + 0\cdot9367} = 0\cdot9393 \quad (۱۱)$$

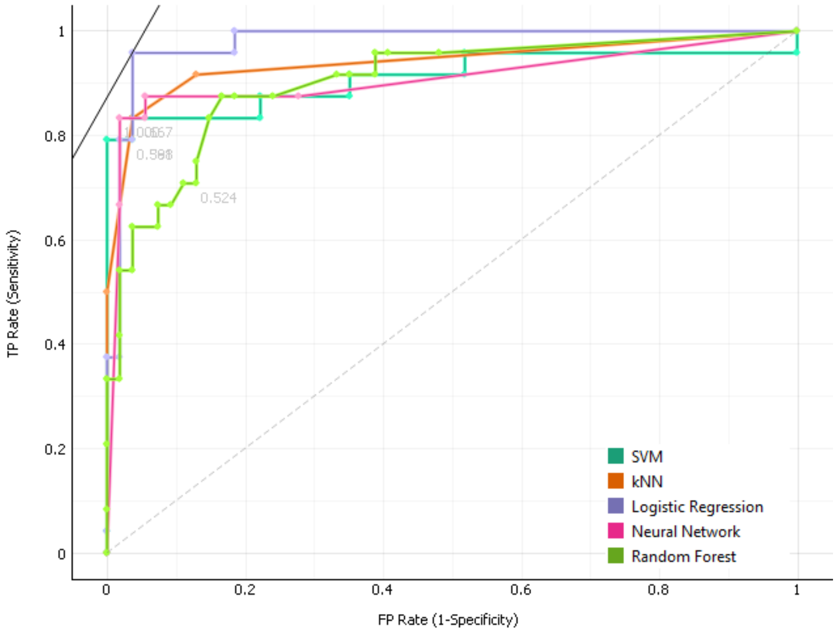


شکل ۵: ماتریس درهم ریختگی مدل SVM در نرم‌افزار Orange

مقایسهٔ مدل‌ها بااستفاده از شاخص AUC در جدول (۳) آورده شده‌است. در این جدول مدل‌ها دوبه‌دو براساس یک شاخص ارزیابی انتخابی با یکدیگر مقایسه می‌شوند. این مقایسه‌ها بااستفاده از تفسیر بیزی از آزمون t است. عددی که در جدول در هر سلول نوشته شده‌است، بیانگر آن است که به احتمال زیاد همان عدد مدل مربوط به سطر سلول از مدل مربوط به ستون سلول بهتر است. جدول (۲) نیز همانند این جدول نتیجه‌گیری مشابه دارند و از هر دو می‌توان یک نتیجه را برداشت کرد. بااستفاده از آنها می‌توان گفت که براساس شاخص AUC روش رگرسیون لاجستیک بهترین روش است؛ اما اگر این شاخص را نادیده بگیریم، براساس سایر شاخص‌ها روش‌های SVM، ANN و رگرسیون لاجستیک عملکردی مشابه هم دارند و از هر سهٔ آنها می‌توان برای پیش‌بینی استفاده کرد. یکی از روش‌های مقایسهٔ عملکرد مدل‌های طبقه‌بندی دودویی (Binary Classifier) استفاده از نمودار مشخصهٔ عملکرد ROC است. ارزیابی عملکرد مدل‌های طبقه‌بندی دودویی معمولا بااستفاده از شاخص‌هایی به نام حساسیت (Sensitivity) و بازیابی (Recall) انجام می‌شود؛ درحالی‌که این نمودار هر دو شاخص را دارا می‌باشد و موجب بررسی هر دوی آنها به‌صورت همزمان می‌شود. در نمودار مشخصهٔ عملکرد بر محور افقی نرخ مثبت کاذب FPR (False Positive Rate) و بر محور عمودی نرخ مثبت صحیح TPR (True Positive Rate) قرار می‌گیرد. در این نمودار یک خط قطری وجود دارد که ناحیهٔ بالای این خط، ناحیهٔ مطلوب و ناحیهٔ زیر آن ناحیهٔ نامطوب محسوب می‌شود. منحنی ترسیم‌شده برای هر مدل طبقه‌بندی هرچه به بالا و سمت چپ نمودار نزدیک‌تر باشد از توانایی و دقت بالاتری برای تشخیص صحیح کلاس داده‌ها برخوردار است. به‌عبارت دیگر می‌توان گفت، روشی که دقت و کارایی بالاتری دارد منحنی آن در نمودار ROC مساحت بیشتری را در زیر خود شامل می‌شود. در نمودار شکل (٦) همان‌طور که مشخص است منحنی مربوط به روش رگرسیون لاجستیک نسبت‌به سایر آنها بالاتر است و پس‌از آن منحنی مربوط به روش KNN قرار می‌گیرد و پایین‌تر از همه روش جنگل تصادفی می‌باشد.

جدول ۳: مقایسه دو به دو مدل‌ها براساس شاخص AUC

مدل‌های طبقه بندی	نزدیک‌ترین همسایگی	ماشین بردار پشتیبان	جنگل تصادفی	شبکه عصبی هوشمند	رگرسیون لاجستیک
نزدیک‌ترین همسایگی	-	۰/٦۱۱	۰/٤۰٤	۰/٥٧۰	۰/۱۹۳
ماشین بردار پشتیبان	۰/۳۸۹	-	۰/۳۰۳	۰/٤۱٧	۰/۱۹۳
جنگل تصادفی	۰/٥۹٦	۰/٦۹٧	-	۰/٦٥۳	۰/۰۸۳
شبکه عصبی هوشمند	۰/٤۳۰	۰/٥۸۳	۰/۳٤٧	-	۰/۲۰۹
رگرسیون لاجستیک	۰/۸۰٧	۰/۸۰٧	۰/۹۱٧	۰/٧۹۱	-



شکل ٦: نمودار مشخصه عملکرد روش‌های طبقه‌بندی هوشمند

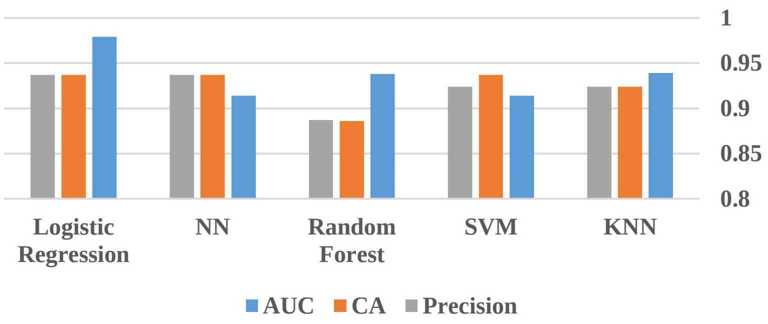
جدول ٤: مقایسه نتایج تحقیقات مختلف در مورد یک مطالعه موردی

محققین	مدل	CA	مرجع
Xue, Yang	ANFIS	۰/۹٦۳	[۳]
Xue, Yang	PSO-SVM	۰/۹٦٥	[۲۱]
تحقیق حاضر	رگرسیون لاجستیک	۰/۹۳۷	-

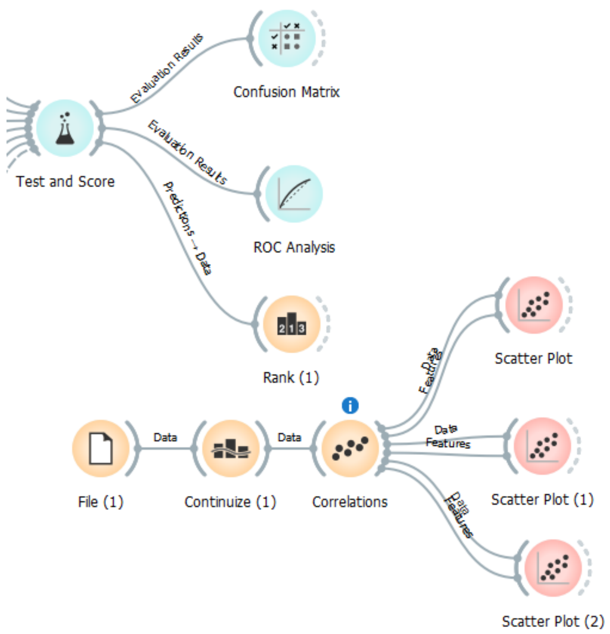
در جدول (٤) نتایج عملکرد سه مدل براساس معیار صحت طبقه‌بندی بیان شده‌است. دو مورد از مدل‌ها در تحقیقات دیگری انجام شده که مدل‌سازی‌ها به‌صورت دستی بوده‌است. در این تحقیق نیز مدل رگرسیون لاجستیک برای مقایسه انتخاب شده‌است. نتایج تحقیق حاضر در مقایسه با دو تحقیق دیگر [3, 21] نشان می‌دهد، نتایج با اختلاف اندکی به هم شبیه هستند؛ ضمن این‌که نرم‌افزار Orange کار مدل‌سازی را بسیار آسان نموده‌است و در کوتاه‌ترین زمان می‌توان انواع مدل‌ها را ساخت و ارزیابی کرد.

در نمودار شکل (۷) سه شاخص AUC، CA و Precision به‌دست‌آمده از هر یک از مدل‌های طبقه‌بندی هوشمند با هم مقایسه شده‌اند. همان‌طور که نشان داده شده‌است روش رگرسیون لاجستیک براساس شاخص AUC با اختلاف زیادی روش برتر محسوب می‌شود؛ اما اگر این شاخص را کنار بگذاریم، روش‌های رگرسیون لاجستیک، شبکهٔ عصبی مصنوعی و ماشین بردار پشتیبان براساس دو شاخص دقت و صحت طبقه‌بندی هوشمند نسبت‌به دو روش دیگر بهتر عمل می‌کنند و می‌توان گفت عملکرد نزدیک به هم دارند. هم‌چنین این نمودار نشان می‌دهد که روش جنگل تصادفی نسبت‌به سایر روش‌ها عملکرد ضعیفی دارد.

شکل (۸) انجام یکی دیگر از قابلیت‌های نرم‌افزار Orange به نام Rank را نشان می‌دهد که به رتبه‌بندی و امتیازدهی متغیرها باتوجه به ارتباط آنها با متغیر هدف می‌پردازد و این کار را براساس روش‌های مختلف ارزش‌گذاری مانند روش‌های بهرهٔ اطلاعاتی، بهرهٔ اطلاعاتی نسبی، شاخص جینی و شاخص ReliefF انجام می‌دهد. خروجی این عملیات در جدول (٥) نشان داده شده‌است. بر این اساس متغیر مقاومت نوک مخروط در آزمایش CPT با اختلاف بسیاری در هر چهار معیار در اولویت اول قرار گرفته‌است. به این معنی که مؤثرترین متغیر در کلاس‌بندی صحیح داده‌ها می‌باشد.



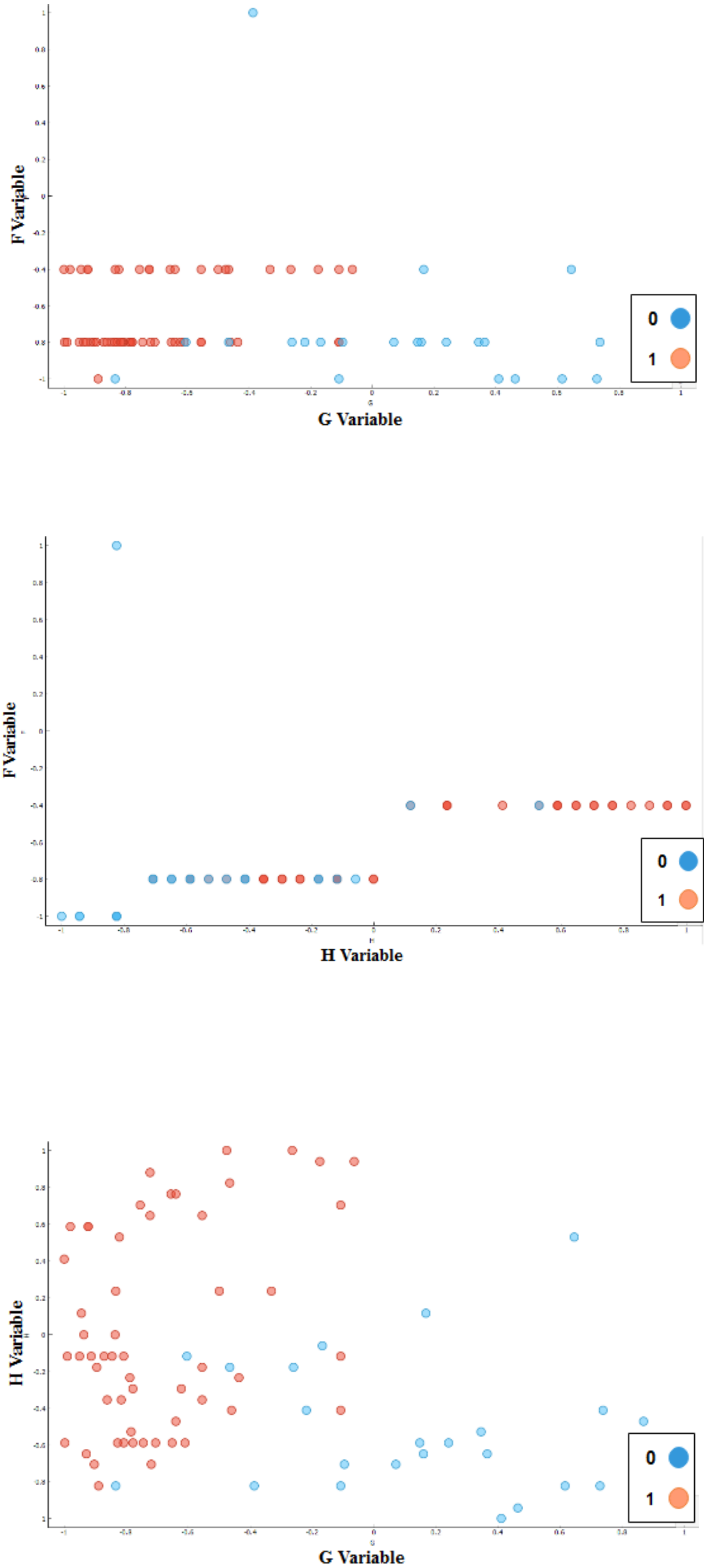
شکل ۷: نمودار ستونی معیارهای ارزیابی روش‌های طبقه‌بندی هوشمند



شکل ۸: استفاده از قابلیت Rank و نمودارهای پراکندگی در نرم افزار Orange

جدول ۵: رتبه‌بندی متغیرهای ورودی براساس روش‌های ارزش‌گذاری

پارامترها	بهره اطلاعاتی	بهره اطلاعاتی نسبی	شاخص جینی	شاخص ReliefF
q_c (MPa)	۰/۴۱۷	۰/۲۰۹	۰/۲۱۶	۰/۲۶۵
τ_{av}/σ'_{v0}	۰/۲۰۸	۰/۱۰۵	۰/۱۱۸	۰/۱۲۴
α_{max} (g)	۰/۱۸۴	۰/۱۳۳	۰/۱۰۶	۰/۰۵۴
d_w (m)	۰/۱۵۸	۰/۰۷۹	۰/۰۸۶	۰/۰۶۶
σ'_{v0}(KPa)	۰/۱۳۱	۰/۰۶۵	۰/۰۷۴	۰/۰۱۶
D_{50} (mm)	۰/۰۸۴	۰/۰۴۲	۰/۰۴۴	۰/۰۲۶
d_s (m)	۰/۰۵۴	۰/۰۲۷	۰/۰۳۱	۰/۰۱۲
σ_v (KPa)	۰/۰۴۸	۰/۰۲۴	۰/۰۲۸	۰/۰۱۳
M	۰	۰	۰	۰



شکل ۹: نمودار پراکندگی متغیرها

برای بررسی سه متغیری که بیشترین تأثیرگذاری را در روند مدل‌سازی داشتند، نمودار پراکندگی آنها در شکل (۹) رسم شده‌است تا تغییرات دوبه‌دوی متغیرها نسبت‌به یکدیگر مشخص شود. همان‌طور که نشان داده شده‌است، بین دو متغیر نسبت تنش تناوبی (H) و حداکثر شتاب در سطح زمین (F) و هم‌چنین دو متغیر مقاومت نوک مخروط (G) و حداکثر شتاب در سطح زمین (F)، هم‌بستگی و نظم خاصی مشاهده نمی‌شود؛ اما بین دو متغیر مقاومت نوک مخروط (G) و نسبت تنش تناوبی (H)، در هر کلاس از روان‌گرایی، یک جهت‌گیری تقریبی وجود دارد، به‌طوری‌که داده‌ها با کلاس ۱ در سمت چپ نمودار و داده‌ها با کلاس صفر در پایین نمودار تجمع یافته‌اند.

نتیجه‌گیری

ارزیابی پتانسیل روان‌گرایی خاک درخصوص پایگاه دادهٔ زلزلهٔ تنگشان با ۵ الگوریتم طبقه‌بندی هوشمند، رگرسیون لاجستیک، SVM، KNN، ANN و جنگل تصادفی و با کمک نرم‌افزار Orange صورت گرفت. مدل‌سازی از ۹ متغیر عددی ورودی و یک متغیر کیفی هدف تشکیل شده بود. نتایج این تحقیق نشان می‌دهد:

- مدل‌های طبقه‌بندی هوشمند انجام‌شده برای پیش‌بینی پتانسیل روان‌گرایی خاک از توانایی و دقت بالایی برخوردار هستند. باتوجه به مقایسه‌هایی که براساس معیارهای مختلف بین روش‌ها انجام شد، روش رگرسیون لاجستیک براساس شاخص AUC بهترین روش معرفی شد.
- براساس شاخص‌های بازیابی، دقت و صحت طبقه‌بندی، سه مدل رگرسیون لاجستیک، SVM و ANN عملکردی نزدیک به هم داشتند و از دقت و صحت بالایی برخوردار بودند. در بین مدل‌سازی‌ها، مدل جنگل تصادفی نسبت‌به سایر مدل‌ها عملکرد ضعیف‌تری داشت.
- بررسی تأثیرگذاری متغیرهای ورودی و ارزش‌گذاری آنها براساس چهار معیار بهرهٔ اطلاعاتی، بهرهٔ اطلاعاتی نسبی، شاخص جینی و شاخص ReliefF نشان داد متغیر مقاومت نوک مخروط در آزمایش CPT اندازه‌گیری شده، نقش عمده‌ای در پیش‌بینی صحیح کلاس روان‌گرایی خاک داشته‌است و پس از آن هم متغیرهای نسبت تنش تناوبی و حداکثر شتاب افقی زلزله در سطح زمین از تأثیرگذاری بالایی در مدل‌سازی‌ها برخوردار هستند.

مراجع

1. Seed, H.B.J.E.e.r.i., "Ground motions and soil liquefaction during earthquakes", *Earthquake engineering research insititue*, Vol. 5, pp. 1249-1273, (1982).
2. Xue, X. and Yang, X.J.N.h., "Application of the adaptive neuro-fuzzy inference system for prediction of soil liquefaction", *Natural hazards*, Vol. 67, pp. 901-917, (2013).
3. Seed, H.B. and Idriss, I.M., "Simplified procedure for evaluating soil liquefaction potential" *Journal of the Soil Mechanics and Foundations division*, Vol. 97, pp. 1249-1273, (1971).
4. Kiang, M.Y.J.D.s.s., "A comparative assessment of classification methods", *Decision support systems*, Vol. 35, pp. 441-454, (2003).
5. Ramakrishnan, D., et al., "Artificial neural network and liquefaction susceptibility assessment: a case study using the 2001 Bhuj earthquake data, Gujarat, India", *Computational Geosciences*, Vol. 12, pp.

491-501, (2008).

6. Chern, S.-G., Lee, C.-Y.J.J.o.M.S., and Technology, "CPT-based simplified liquefaction assessment by using fuzzy-neural network", *Journal of Marine Science and Technology*, Vol. 17, pp. 326-331, (2009).
7. Mughieda, O., Bani-Hani, K., and Safieh, B.J.I.J.o.G.E., "Liquefaction assessment by artificial neural networks based on CPT", *International Journal of Geotechnical Engineering*, Vol. 3, pp. 289-302, (2009).
8. Sulewska, M.J.J.C.A.M.i.E. and Science, "Applying artificial neural networks for analysis of geotechnical problems", *Computer Assisted Methods in Engineering and Science*, Vol. 18, pp. 231-241, (2017).
9. Samui, P., Sitharam, T.J.N.H., and Sciences, E.S., "Machine learning modelling for predicting soil liquefaction susceptibility", *Natural Hazards and Earth System Sciences*, Vol. 11, pp. 1-9, (2011).
10. Farrokhzad, F., Choobbasti, A., and Barari, A.J.J.o.K.S.U.-S., "Liquefaction microzonation of Babol city using artificial neural network", *Journal of King Saud University*, Vol. 24, pp. 89-100, (2012).
11. Tolon, M.J.I.J.H.S., "A comparative study on computer aided liquefaction analysis methods", *Int. Journal for Housing Science*, Vol. 37, pp. 121-35, (2013).
12. Muduli, P.K. and Das, S.K.J.I.G.J., "CPT-based seismic liquefaction potential evaluation using multi-gene genetic programming approach", *Indian Geotechnical Journal*, Vol. 44, pp. 86-93, (2014).
13 .Bre, F., et al., "Prediction of wind pressure coefficients on building surfaces using artificial neural networks", *Energy and Buildings*, Vol. 158, pp. 1429-1441, (2018).
14. Mashrei, M.A.J.F.I.S.-T.," Neural network and adaptive neuro-fuzzy inference system applied to civil engineering problems", *Fuzzy Inference System-Theory and Applications*, Vol., (2012).
15. Noble, W.S.J.N.b., "What is a support vector machine?", *Nature biotechnology*, Vol. 24, pp. 1565-1567, (2006).
16. Wu, X. and Kumar, V., "*The top ten algorithms in data mining"*. CRC pres, (2009).
17. Wright, R.E., "Logistic regression", *American Psychological Association*, Vol., (1995).
18. Breiman, L.J.M.l., "Random forests", *Machine learning*, Vol. 45, pp. 5-32, (2001).
19. Raileanu, L.E., Stoffel, K.J.A.o.M., and Intelligence, A., "Theoretical comparison between the gini index and information gain criteria", *Annals of Mathematics and Artificial Intelligence*, Vol. 41, pp. 77-93, (2004).
20. Kira, K. and Rendell, L.A. The feature selection problem: Traditional methods and a new algorithm. in Aaai. 1992.
21. Xue, X., Yang, X.J.B.o.E.G., and Environment, t., "Seismic liquefaction potential assessed by support vector machines approaches", *Bulletin of Engineering Geology and the Environment*, Vol. 75, pp. 153-162, (2016).